Министерство образования Республики Беларусь

Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК 004.04:339.16

Трубчик
Артём Иванович

СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ В ЗАДАЧАХ ЭЛЕКТРОННОЙ ТОРГОВЛИ

АВТОРЕФЕРАТ

на соискание степени магистра технических наук

по специальности 1-40 80 03 «Вычислительные машины и системы»

_____

Научный руководитель
Самаль Дмитрий Иванович
кандидат технических наук, доцент

_____

Минск 2016

КРАТКОЕ ВВЕДЕНИЕ

Для достижения успеха в электронной торговле, очень важен анализ рынка. Существует два направления анализа: технический и фундаментальный. В техническом анализе, используется информация о биржевой торговле на фондовом рынке, такая как цена акции и объем торгов для определения будущей цены акции, в то время как фундаментальный анализ оперирует информацией, полученной за пределами фондового рынка, в частности, операционной прибылью, экономической ситуацией, тенденциями эксплуатационных затрат для прогнозирования движений цен акций.

Теоретическим основанием значительной части критики финансового прогнозирования является гипотеза эффективного рынка. Согласно этой гипотезе, цены на финансовых рынках точно отражают всю имеющуюся на данный момент информацию и следовательно, стабильная генерация прибыли выше средней по рынку невозможна. Но в то же время важным выводом из гипотезы эффективного рынка является то, что если удастся собрать некую информацию, не входящую в рынок, то она может быть использована для получения экономического преимущества.

Одним из способов получения новой информации является интеллектуальных анализ данных. В рамках этой работы исходными данными для анализа выступают сообщения социальной сети Twitter и торговая история криптографической валюты Bitcoin. Подобные исследования использовали анализ тональности текстов Twitter и проводились только применительно к фондовым биржам, прогнозированию популярности кинофильмов и оценки общественного мнения.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Целью этой работы является создание системы интеллектуального анализа данных для повышения результативности торговли на рынке криптографической валюты Bitcoin. Для этого необходимо было решить следующие задачи:

– загрузить, очистить и подготовить данные Twitter и Bitcoin;

– создать алгоритм оценки тональности сообщений Twitter;

– исследовать взаимосвязь между данными из твитов и торговой историей Bitcoin;

– спрогнозировать движение котировок Bitcoin не только на основе торговой истории, но и данных, полученных из сообщений Twitter.

Объект исследования: рынок криптографической валюты Bitcoin.

Предмет исследования: методы анализа открытых информационных источников, таких как Twitter.

В работе предложен простой и достаточно эффективный алгоритм определения тональности сообщений Twitter. Впервые проведен корреляционный анализ между данными, полученных из сообщений Twitter и торговой истории Bitcoin. Представлены результаты прогнозирования движения котировок рынка Bitcoin. Итогом работы является разработанная система интеллектуального анализа данных.

Основные положения и результаты работы нашли отражение в 2 публикациях автора.

Цели и задачи работы обуславливают ее структуру, которая состоит из введения, пяти глав основной части, заключения, списка использованных источников и приложений. В начальной главе существляется краткий обзор проблемы, в последующих главах предложенные методы подготовки данных и анализа приводятся совместно с результатами исследований. В конце каждой главы представлены выводы по всем результатам.

Диссертация выполнена на 82 страницах, включая 2 приложения информационного характера. Включает в себя 5 глав, 22 иллюстрации, 12 таблиц, 19 формул, 43 библиографических источника.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В данном списке перечислены основные результаты по главам:

– В главе 1 был проведен обзор проблемы, описана гипотеза эффективного рынка и установлены различные подходы к анализу тональности текста, которые включают в себя классификаторы и лексические подходы. Подробно изучен ряд работ посвященных прогнозированию на основе публичной информации.

– В главе 2 был предложен и реализован метод для оценивания тональности текста в коротких сообщениях на основе лексического анализа. Представлено несколько примеров, в которых публикация важных новостей привела к значительному увеличению объема сообщений по теме Bitcoin.

– В главе 3 приведено несколько значимых корреляций между данными Twitter и рыночными данными Bitcoin, или между их производными. Примеры включают несколько случаев, когда значительные колебания индикатора Twitter предшествует значительным колебаниям рыночных показателей. Эти результаты дают основание говорить о том, что данные, полученные из Twitter, могут содержать информацию, которая может быть использована для прогнозирования изменений рынка.

– В главе 4 было применено машинное обучение в виде метода опорных векторов, описаны используемые преобразования и методика оценки результатов. Установлена точность прогноза без применения публичной информации в $56{,}49\%$, с применением в $63{,}27\%$.

– В главе 5 представлена разработанная система интеллектуального анализа данных в реальном времени. Создана на языке Scala с использованием фреймворка Apache Spark и подсистемы визуализации Apache Zeppelin. Предложен метод оценки качества сообщений для создания списка самых интересных новостей за 24 часа.

ЗАКЛЮЧЕНИЕ

В главах 3-4 благодаря массовому корреляционному анализу и эксперименту по прогнозированию фактически было установлено наличие взаимосвязи между данными, которые можно извлечь из социальной сети Twitter и торговой истории Bitcoin.

Используемые методы подготовки данных, в частности анализ тональности текста, позволил достичь точности прогноза в $63{,}27\%$. К сожалению, этого недостаточно, чтобы получить преимущество перед другими трейдерами на торговых площадках Bitcoin. Последние вычитают комиссию за каждый исполненный ордер в $0{,}1\%$ (применяется дважды для обеих сторон сделки, то есть к каждому из ордеров на покупку и продажу), что полностью нивелирует полученное преимущество в виде прогнозирования движения котировок рынка.

В дальнейшем следует продолжать работу над подходами к очистке и подготовке данных. Например, в исходном наборе данных, представленном в главе 2, используются только те сообщения, которые содержат слова «bitcoin», «btc» и «биткоин». Вполне возможно, что объема сообщений по этим ключевым словами недостаточно для точного анализа и прогнозирования рынка и одним из вариантов усовершенствования прогноза может быть расширение базы поступающих твитов с помощью новых ключевых слов до максимального ограничения, установленного Twitter. Безусловно, это увеличит размер базы данных в несколько раз, но масштабируемость системы, описанная в главе 5, позволяет это сделать.

Примененные методы и разработанная система анализа может быть использованы не только в задачах электронной торговли. Если заменить часть системы, связанную с обработкой и прогнозированием торговой истории на предсказывание каких-либо событий в мире, то мы получим новую систему анализа, использующую данные Twitter, но применительно к новым знаниям. Например, это может быть оценка популярности бренда, прогноз кассовых сборов только выпущенного в прокат фильма, прогноз президентских выборов или просто анализ настроений пользователей по какому-либо ключевому слову.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

Трубчик, А.И. Twitter как индикатор в задачах электронной торговли / А.И. Трубчик // 51-я научная конференция аспирантов, магистрантов и студентов по направлению 4: Компьютерные системы и сети – Минск : БГУИР, 2015. – С. 22.

Трубчик, А.И. Оценка качества сообщений Twitter для интеллектуального анализа данных / А.И. Трубчик // Информационные технологии и системы 2015 (ИТС 2015) : материалы международной научной конференции (БГУИР, Минск, Беларусь, 28 октября 2015) – Минск : БГУИР, 2015. – С. 310-311